Understanding the Generalization Benefit of Model Invariance from a Data Perspective

Sicheng Zhu*, Bang An*, Furong Huang
Department of Computer Science
University of Maryland, College Park
{sczhu, bangan, furongh}@umd.edu

Abstract

Machine learning models that are developed to be invariant under certain types of data transformations have shown improved generalization in practice. However, a principled understanding of why invariance benefits generalization is limited. Given a dataset, there is often no principled way to select "suitable" data transformations under which model invariance guarantees better generalization. This paper studies the generalization benefit of model invariance by introducing the *sample cover induced by transformations*, i.e., a representative subset of a dataset that can approximately recover the whole dataset using transformations. For any data transformations, we provide refined generalization bounds for invariant models based on the sample cover. We also characterize the "suitability" of a set of data transformations by the *sample covering number induced by transformations*, i.e., the smallest size of its induced sample covers. We show that we may tighten the generalization bounds for "suitable" transformations that have a small sample covering number. In addition, our proposed sample covering number can be empirically evaluated and thus provides a guidance for selecting transformations to develop model invariance for better generalization. In experiments on multiple datasets, we evaluate sample covering numbers for some commonly used transformations and show that the smaller sample covering number for a set of transformations (e.g., the 3D-view transformation) indicates a smaller gap between the test and training error for invariant models, which verifies our propositions.

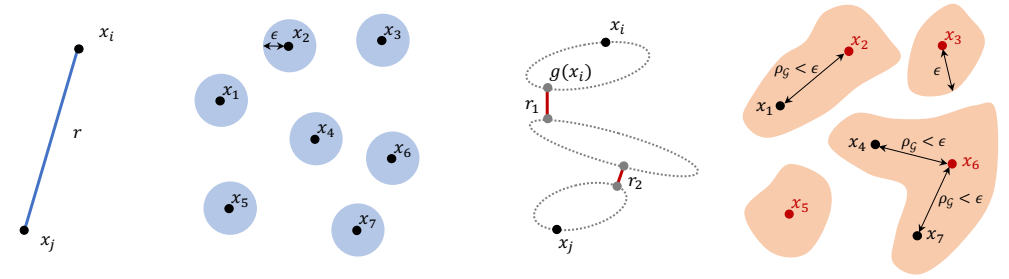
1 Introduction

Invariance is ubiquitous in many real-world problems. For instance, categorical classification of visual objects is invariant to slight viewpoint changes [18, 2, 23], text understanding is invariant to synonymous substitution and minor typos [49, 34, 25]. Intuitively, models capturing the underlying invariance exhibit improved generalization in practice [21, 13, 46, 14, 42, 14]. Such generalization benefit is especially crucial when the data are scarce as in some medical tasks [43], or when the task requires more data than usual as in cases of distribution shift [35] and adversarial attack [37, 45, 5].

A commonly accepted intuition attributes the generalization benefit of model invariance to the reduced model complexity, especially the reduced sensitivity to spurious features. However, a principled understanding of why model invariance helps generalization remains elusive, thus leaving many open questions. Since model invariance may come at a cost (e.g., compromised accuracy, increase computational overhead), given a task, how should we choose among various data transformations under which model invariance guarantees better generalization? If existing data transformations are not good enough for a given task, what is the guiding principle to find new ones? The lack of a principled understanding limits better exploitation of model invariance to further improve

* Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).



$\rho_{\mathcal{G}}(x_i, x_j) = r$ sample cover $= \{x_1, x_2, \ldots, x_7\}$

(a) $\mathcal{G} = \emptyset$

$\rho_{\mathcal{G}}(x_i, x_j) = r_1 + r_2$ sample cover $= \{x_2, x_3, x_5, x_6\}$

(b) $\mathcal{G} = \{$"rotation"$\}$

Figure 1: Illustration of the pseudometric and sample cover induced by data transformations.

generalization. In addition, since identifying instructive generalization bound is a central topic in machine learning, we may expect to tighten existing generalization bounds by additionally considering the data-dependent model invariance property.

The many faces of data transformations and model classes pose significant challenges to a principled understanding of model invariance's generalization benefit. To address this, [2, 4, 32, 39, 36] characterize the input space and show that certain data transformations equivalently shrink the input space for invariant models, which then simplify the input and improves generalization. From another perspective, [19, 28] directly characterize the function space and show that the volume of the invariant model class is reduced, which then simplifies the learning problem and improves generalization. These understandings provide valuable insights, yet they may become less informative on high-dimensional input data or require the model invariance to be obtained exclusively via feature averaging. Some certain assumptions on data transformations (e.g., finiteness, group structure with certain measure) also make these understandings less applicable to more general data transformations.

In this paper, we derive generalization bounds for invariant models based on the sample cover induced by data transformations and empirically show that the introduced notion can guide the data transformations selection. Different from previous understandings, we first identify a data-dependent property of data transformations in a model-agnostic way, and then establish its connections with the refined generalization bounds of invariant models. The analysis applies to more general data transformations regardless of how the model invariance is obtained and naturally provides a model-agnostic guidance for data transformations selection. We summarize our contributions as follows.

Contributions. At the core of our understanding is the notion of sample cover induced by data transformations, defined informally as a representative subset of a dataset that can approximately recover the whole dataset using data transformations (illustrated in Figure 1). We show that this notion identifies a data-dependent property of data transformations which is related to the generalization benefit of the corresponding invariant models. Under a special setting of the sample cover, we first bound the model complexity of any invariant and output-bounded model class in terms of the sample covering numbers. Since this general bound requires a restrictive condition on data transformations in order to be informative, we then assume the model Lipschitzness to relax the requirement and refine the model complexity bound for invariant models. Finally, we outline a framework for model-invariance-sensitive generalization bounds based on the invariant models' complexities, and use it to discuss the generalization benefit of model invariance.

Given the usefulness of sample cover in the analysis, we propose an algorithm to empirically estimate the sample cover. This algorithm exactly verifies whether a given subset of a sample forms a valid sample cover, and always estimates a sample covering number that upper-bounds the ground-truth. Inspired by our analysis, we also propose to use the sample covering number as a suitability measurement for practical data transformation selections. This measurement is data-driven, widely applicable, and empirically correlates with invariant models' actual generalization performance. We discuss its limitations and the empirical mitigation.

To empirically verify our propositions, we first estimate the sample covering number for some commonly used data transformations on four image datasets, including CIFAR-10 and ShapeNet (a 3D dataset). Under typical settings, the 3D-view transformation induces a much smaller sample covering number than others on ShapeNet, while cropping induces the smallest sample covering number on others datasets. For those data transformations, we then train invariant models via data augmentation and invariance loss regularization to evaluate the actual generalization benefit. Results show a clear correlation between smaller sample covering numbers induced by data transformations and the better generalization benefit enjoyed by invariant models.

2 Preliminaries

Data transformations. We refer to the data transformation as a function from the input space $\mathcal{X} \to \mathcal{X}$, and data transformation*s* as a set of such functions. Unless otherwise specified, we do not assume data transformations to have group structures since many non-invertible transformations (e.g., cropping) do not fit into a group structure directly. For a set of data transformations $\mathcal{G} = \{g : \mathcal{X} \to \mathcal{X}\}$ and a data point (also referred to as an example) $\boldsymbol{x} \in \mathcal{X}$, we overload the notion of orbit in group theory and denote by $\mathcal{G}(\boldsymbol{x})$ the orbit of $\boldsymbol{x}$ defined as follows. The *orbit* of $\boldsymbol{x}$ generated by data transformations $\mathcal{G}$ is the collection of outputs after applying any transformation $g \in \mathcal{G}$ on $\boldsymbol{x}$: $\mathcal{G}(\boldsymbol{x}) = \{g(\boldsymbol{x}) \in \mathcal{X} : g \in \mathcal{G}\}$.

Model invariance. Let $\mathcal{D}$ be the underlying data distribution and $\text{supp}(\mathcal{D})$ be its support. A model $h : \mathcal{X} \to \mathcal{Y}$ is said to be *invariant* under data transformations $\mathcal{G}$ on $\mathcal{D}$ if $h(g(\boldsymbol{x})) = h(\boldsymbol{x})$ for any $\boldsymbol{x} \in \text{supp}(\mathcal{D})$ and any $g \in \mathcal{G}$. We refer to a class of invariant models as the *$\mathcal{G}$-invariant model class*.

Complexity measurements. *Covering number* and *Rademacher complexity* [31] are two commonly used complexity measurements for model classes (including neural networks [6]) that can provide uniform generalization bounds. The covering number can also be directly used to upper bound the Rademacher complexity via Dudley's entropy integral theorem [17, 30].

Covering number. Let $(\mathcal{F}, d)$ be a (pseudo)metric space with some (pseudo)metric[1] d. An *ϵ-cover* of a set $\mathcal{H} \subseteq \mathcal{F}$ is defined as a subset $\widehat{\mathcal{H}} \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$, there exists $\widehat{h} \in \widehat{\mathcal{H}}$ such that $d(h, \widehat{h}) \leq \epsilon$. The covering number $N(\epsilon, \mathcal{H}, d)$ is defined as the minimum cardinality of an ϵ-cover (among all ϵ-covers) of $\mathcal{H}$. In this paper, we use the concept of covering number both for measuring model class complexities and for defining the sample covering number on datasets.

Empirical Rademacher complexity. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathbb{R}$. Given a sample $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, the *empirical Rademacher complexity* of model class $\mathcal{H}$ is defined as: $\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right]$ where $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_n]^\top$ is the vector of i.i.d. Rademacher random variables, each uniformly chosen from $\{-1, 1\}$.

Generalization error and gap. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample drawn i.i.d. from some data distribution $\mathcal{D}$, and $\mathcal{H}$ be a model class. Given a loss function $\ell : \mathbb{R} \to [0, 1]$, for a $h \in \mathcal{H}$, we define the *empirical error* as $R_{\mathcal{S}}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(\boldsymbol{x}_i), y_i)$, the *generalization error* as $R(h) = \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}}[\ell(h(\boldsymbol{x}), y)]$, and the *generalization gap* as $R(h) - R_{\mathcal{S}}(h)$.

3 Generalization Benefit of Model Invariance

In this section, we derive the generalization bounds for invariant models by identifying the model invariance properties. We start by introducing the notion of sample cover induced by data transformations and based on it bound the Rademacher complexity of any invariant models with bounded output (Section 3.1). Then, we assume model Lipschitzness to provide a more informative model complexity bound for any data transformations (Section 3.2). Finally, we provide a framework for model-invariance-sensitive generalization bounds and discuss the generalization benefit of model invariance (Section 3.3).

[1] A pseudometric is a metric if and only if it separates distinct points, namely $d(x, y) > 0$ for any $x \neq y$.

3.1 Sample Cover Induced by Data Transformations

Existing empirical results suggest that, compared with standard models, invariant models may have certain properties reducing their effective model complexities. To identify such properties, we alternatively identify the related properties of the corresponding data transformations via the notion of *sample cover induced by data transformations*. We now formalize the introduced notion.

The definition of sample cover relies on the pseudometric induced by the data transformations $\mathcal{G}$. Note that $\mathcal{G}$ generates an orbit $\mathcal{G}(\boldsymbol{x}) \subseteq \mathcal{X}$ for each example $\boldsymbol{x} \in \mathcal{S}$. Let $\|\cdot\|$ be any norm on the input space $\mathcal{X}$. Given a set of transformations $\mathcal{G}$, we define the $\mathcal{G}$-induced pseudometric[2] as

$$\rho_{\mathcal{G}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \inf_{\gamma \in \Gamma(\boldsymbol{x}_1, \boldsymbol{x}_2)} \int_{\gamma} c(\boldsymbol{r}) ds, \quad \text{where } c(\boldsymbol{r}) = \begin{cases} 0, & \text{if } \boldsymbol{r} \in \cup_{\boldsymbol{x} \in \mathcal{S}} \mathcal{G}(\boldsymbol{x}) \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

where $ds = \|d\boldsymbol{r}\|$, and Γ denotes the set of all paths (curves) in $\mathcal{X}$ from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$. The $\rho_{\mathcal{G}}$ is essentially calculating the line integral along the shortest (if achievable) path γ in the scalar field c, where c can also be viewed as the "moving cost" function depending on $\mathcal{G}$. The norm $\|\cdot\|$ here can be selected as any meaningful norm on the input space (e.g., Euclidean norm as in our experiments) and will later be used in defining the model's Lipschitz constant. It can be checked that $\rho_{\mathcal{G}}$ satisfies pseudometric axioms.

Definition 3.1 (Sample cover induced by data transformations)**.** Let $(\mathcal{X}, \rho_{\mathcal{G}})$ be a pseudometric space and $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. An *ϵ-sample cover* $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ of the sample $\mathcal{S}$ induced by data transformations $\mathcal{G}$ at resolution ϵ is defined as a subset of the sample $\mathcal{S}$ such that for any $\boldsymbol{x} \in \mathcal{S}$, there exists $\widehat{\boldsymbol{x}} \in \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ satisfying $\rho_{\mathcal{G}}(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \leq \epsilon$.

Definition 3.2 (Sample covering number induced by data transformations)**.** The *sample covering number* $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by data transformations $\mathcal{G}$ is defined as the minimum cardinality of an ϵ-sample cover:

$$N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}}) = \min\{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}| : \widehat{\mathcal{S}}_{\mathcal{G},\epsilon} \text{ is an } \epsilon\text{-sample cover of } \mathcal{S}\}. \tag{3.2}$$

Informally, the $\mathcal{G}$-induced sample cover specifies a representative subset of examples which can approximately recover all the original examples using the given data transformations $\mathcal{G}$. This notion is closely related to the *sample compression* [20] which represents a scheme to prove the learnability of concepts through a compressed set of sample. While identifying the generalization-related properties of data transformations, this notion is insensitive to other unrelated properties (e.g., finiteness, group structures) and thus applies to any data transformations.

The intuition behind sample cover is that $\mathcal{G}$-invariant models may have consistent behaviors on an example and its associated approximation in the $\mathcal{G}$-induced sample cover. As such, we can analyze the model complexities of invariant models by considering the models' behaviour only on the potentially small-sized sample covers. Indeed, we directly have the following model complexity result. The proof is in Appendix B.

Proposition 3.3. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}$ be a model class mapping from $\mathcal{X}$ to $[-B, B]$ for some $B > 0$ and is invariant to data transformations $\mathcal{G}$. Then the empirical Rademacher complexity of $\mathcal{H}$ satisfy

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) \leq 24B\sqrt{\frac{N(0, \mathcal{S}, \rho_{\mathcal{G}})}{n}}. \tag{3.3}$$

Proposition 3.3 generally bounds the model complexity of any output-bounded and $\mathcal{G}$-invariant model class in terms of the sample covering number $N(0, \mathcal{S}, \rho_{\mathcal{G}})$ induced by $\mathcal{G}$. A small $\mathcal{G}$-induced sample covering number at resolution $\epsilon = 0$ thus tightens the model complexity bound for a general class of $\mathcal{G}$-invariant models.

Note, however, that Proposition 3.3 is informative only when the data transformations $\mathcal{G}$ yields $N(0, \mathcal{S}, \rho_{\mathcal{G}}) \ll n$ on the sample $\mathcal{S}$ — a condition requiring $\mathcal{G}$ to be able to exactly recover $\mathcal{S}$ from a small-sized subset of $\mathcal{S}$. This condition is unfortunately too strict to hold for many commonly used

[2]Note that $\rho_{\mathcal{G}}$ is not a metric since it allows $\rho_{\mathcal{G}}(x, y) = 0$ for $x \neq y$.

data transformations which only generate orbits with measure zero (with respect to the data measure) at most data points. For example, the rotation transformations on CIFAR-10 do not satisfy this condition, since no two images in CIFAR-10 are rotated versions of each other. To better understand the generalization benefit brought by any data transformations (e.g., rotation), we further assume specific model properties which equivalently expand the orbits in order to get more general results. We study Lipschitz models in Section 3.2, and relegate a sharper (and relatively independent) analysis for linear models under linear data transformations to Appendix C.

3.2 Refined Complexity Analysis of Lipschitz Models

This subsection refines the model complexity analysis for Lipschitz models that are invariant. Characterizing the Lipschitz constant of models has been the focus of a line of work. For example, the Lipschitz constant of ReLU networks can be upper-bounded by the product of the spectral norms of the weight matrices, considering the worst-case inputs [6, 22]. Assuming Lipschitzness, the following theorem refines the covering number analysis for invariant models. The proof is in Appendix B.

Theorem 3.4. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}$ be a model class such that every $h \in \mathcal{H}$ is κ-Lipschitz with respect to $\|\cdot\|$ (used in defining the sample cover) and is invariant to $\mathcal{G}$. Then the covering number of $\mathcal{H}$ satisfies

$$N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\mathcal{S}})\big) \leq \inf_{\epsilon \geq 0, \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}} N\Big(\tau - \kappa\epsilon\sqrt{1 - \frac{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}|}{n}}, \mathcal{H}, L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})\Big), \tag{3.4}$$

where $\forall h, g \in \mathcal{H}$, the $L_2(\mathbb{P}_{\mathcal{S}})$ metric is defined as $\|h - g\|_{L_2(\mathbb{P}_{\mathcal{S}})} = \Big(\sum_{\boldsymbol{x}\in\mathcal{S}} \frac{1}{n}\big(h(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2\Big)^{\frac{1}{2}}$, and the $L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})$ metric is defined as[3] $\|h - g\|_{L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})} = \Big(\sum_{\boldsymbol{x}\in\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}} \frac{p(\boldsymbol{x})}{n}\big(h(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2\Big)^{\frac{1}{2}}$.

Theorem 3.4 upper-bounds the covering number of $\mathcal{H}$ evaluated at the sample $\mathcal{S}$ by the new covering number evaluated at any sample cover $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$, under a modified metric and at the cost of an additional error term depending on ϵ and κ. The equality trivially holds by taking $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon} = \mathcal{S}$, while by searching over all sample covers with different resolution ϵ it is possible to tighten the covering number bound for invariant models. Additionally, Theorem 3.4 leads to a refined version of Dudley's entropy integral theorem (see Lemma B.1) that bounds the Rademacher complexity of invariant models.

Theorem 3.4 suggests that we may improve existing covering-number-based model complexity analysis by weakening the dependence on input dimensions. Note that covering numbers that do not yield $N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\mathcal{S}})\big)/n \to 0$ as $n \to \infty$ are vacuous. Therefore, existing covering number results typically avoid linear dependence on n at the cost of (explicitly or implicitly) increased dependence on the input dimension [48]. With the refined result in Theorem 3.4, however, a covering number linear in n can now be replaced by one that is linear in a potentially much smaller sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ and consequently become informative, thus circumvent the increased dependence on input dimensions. An interesting direction for future work is to instantiate the result in Equation 3.4 for specific model classes to get more interpretable results.

3.3 Framework for Model-invariance-sensitive Generalization Bounds

This subsection presents the framework for generalization bounds sensitive to the model invariance. While the results are straightforward applications of the derived complexities of invariant models, our goal is to justify the selection of suitable data transformations to maximize the generalization benefit. We start with the generalization analysis of invariant models and then present the framework.

Generalization benefit for invariant models. The generalization bounds of invariant models follow immediately by applying the Rademacher model complexities (Proposition 3.3, Proposition B.1, and Theorem C.1) to the standard generalization bound (Theorem A.2). Compared with standard models, invariant models' tightened model complexity bounds already imply their reduced generalization gaps,

[3] The term $p(\boldsymbol{x})/n$ can be viewed as the probability mass at $\boldsymbol{x}$ where the numerator indicates the number of examples that $\boldsymbol{x}$ covers. See Appendix B.1 for the formal definition of $p(\boldsymbol{x})$.

whereas for reduced generalization error they further need to have low empirical error. Since enforcing the model invariance may simultaneously increase the empirical error, we can use standard model selection techniques (e.g., structural risk minimization [31]) to select suitable data transformations and control the trade-off.

Model-invariance-sensitive generalization bound. We outline the generalization bound that identifies model invariance properties based on the derived invariant models' complexities. It follows by the post-hoc analysis which specifies a proper set of invariant models using the "invariant loss" — the loss when composed with any model, makes the composition invariant. For data transformations with group structures, we can construct such loss by averaging (assuming Haar measure) or adversarially perturbing any given loss over the orbits of input examples [28, 19]. Specifically, the adversarial loss with respect to data transformations $\mathcal{G}$ is defined as $\widetilde{\ell}_{\mathcal{G}}\left(h(x), y\right) = \max_{x' \in \mathcal{G}(x)} \ell\left(h(x'), y\right)$, where ℓ is any given loss. Using the adversarial loss, the following proposition provides the model-invariance-dependent generalization bound by applying the model selection framework [31]. Appendix B.3 further describes a binary coding construction of combinations of data transformation classes.

Proposition 3.5. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}$ be any given model class and ℓ be any given loss. Suppose we have K sets of group-structured data transformations $\{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_K\}$. Then the following generalization bound holds for any $h \in \mathcal{H}$ and any $k \in [K]$:

$$R(h) \leq \frac{1}{n}\sum_{i=1}^{n} \widetilde{\ell}_{\mathcal{G}_k}(h(\boldsymbol{x}_i), y_i) + 4\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{G}_k} \circ \mathcal{H}) + \sqrt{\frac{\log k}{n}} + 3\sqrt{\frac{\log \frac{4}{\delta}}{2n}}, \tag{3.5}$$

where $\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{G}_k} \circ \mathcal{H})$ is upper-bounded by the complexity of $\mathcal{G}_k$-invariant models. For any model trained on $\mathcal{S}$, Proposition 3.5 shows that we can optimize over all selections of data transformations to improve its generalization bound. Note that the selection of $\mathcal{G}_k$ is subject to a potential trade-off between the reduced model complexity $\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{G}_k} \circ \mathcal{H})$ and the increased empirical error $\sum_{i=1}^n \widetilde{\ell}_{\mathcal{G}_k}(h(x_i), y_i)$. Thus, if a suitable $\mathcal{G}_k$ reduces the model complexity while keeping the empirical error low, then the trained model will benefit from a tightened generalization bound. This generalization bound does not require the models to be (strictly) invariant and potentially explains the improved generalization of models with trained invariance (e.g., via data augmentation [40, 38] or consistency regularization [29, 44]). The difficulty in instantiating Proposition 3.5 is that the model complexity with adversarial loss may be hard to compute for general data transformations. Therefore, we discuss more practical data transformations selections based on the sample covering numbers in Section 5.

4 Sample Cover Estimation Algorithm

The sample cover induced by data transformations plays a central role in our understanding of model invariance. Despite the usefulness in the analysis, exactly computing the sample cover turns out to be non-trivial in general. Indeed, computing the transformation-induced metrics can be difficult for continuous data transformations, and finding the *smallest* sample cover is NP-hard. To address this problem, we propose an algorithm to estimate the sample covering number and find the associated sample cover. We outline the algorithm and discuss the algorithmic challenges in this section. The algorithmic details appear in Appendix D.

Setup. The estimation algorithm takes as input a sample $\mathcal{S}$, a set of data transformations $\mathcal{G}$, and the resolution parameter ϵ. It then returns the estimated sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ and the associated sample cover $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$. The estimation algorithm has the following steps.

Step 1. Compute (or approximate) the direct orbit distance between any two examples in $\mathcal{S}$. The direct orbit distance between any two examples $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}$ is

$$d_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\mathcal{G}(\boldsymbol{x}_i) - \mathcal{G}(\boldsymbol{x}_j)\| = \min_{g_1, g_2 \in \mathcal{G}} \|g_1(\boldsymbol{x}_i) - g_2(\boldsymbol{x}_j)\|,$$

which can be exactly computed for finite transformations (e.g., flipping) with complexity $O(|\mathcal{G}|^2)$), or can be approximated for continuous transformations (e.g., rotation) via optimization or sampling.

Step 2. Compute the $\rho_{\mathcal{G}}$ distance between any two examples in $\mathcal{S}$. Given results in step 1, computing the $\rho_{\mathcal{G}}$ distance between any two examples can be formulated as a shortest path problem on a complete

graph, where each node represents an example and the cost of each edge is the direct orbit distance computed in step 1 (see formulations in Appendix D). Note that the shortest path is always included in our finite candidates even though the $\rho_{\mathcal{G}}$ distance considers infinitely many paths. This is because any other path outside our finite candidates will be longer than its counterparts (depending on what orbits it intersects) in our finite candidates. Standard shortest path algorithms solve for all pairs of examples in polynomial time (e.g., via Dijkstra's algorithm [16] in $O(n^3)$).

Step 3. Construct the pairwise distance matrix $[\rho_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j}$ and approximate the sample covering number. This step can be formulated as a set cover problem where each example $\boldsymbol{x}$ covers a subset of $\mathcal{S}$ in which each element's $\rho_{\mathcal{G}}$ distance to $\boldsymbol{x}$ is less than or equal to ϵ. Our goal is to find a minimum number of those subset such that their union contains $\mathcal{S}$. This problem is known to be NP-hard but admits polynomial time approximations. In experiments, we use modified k-medoids [33] clustering method to find the approximation of $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ (see Algorithm 1).

Note that the estimated sample covering number returned by the algorithm is always an upper bound of the ground-truth, regardless of the approximation error in step 1 and 3. When step 1 is exact, the algorithm also exactly verifies whether a given subset of $\mathcal{S}$ forms a valid sample cover. In our experiment, the step 2 becomes the computation bottleneck for large-sized sample. We leave improving the scalability as well as evaluating the approximation quality for future work.

5 Data-driven Selection of Data Transformations

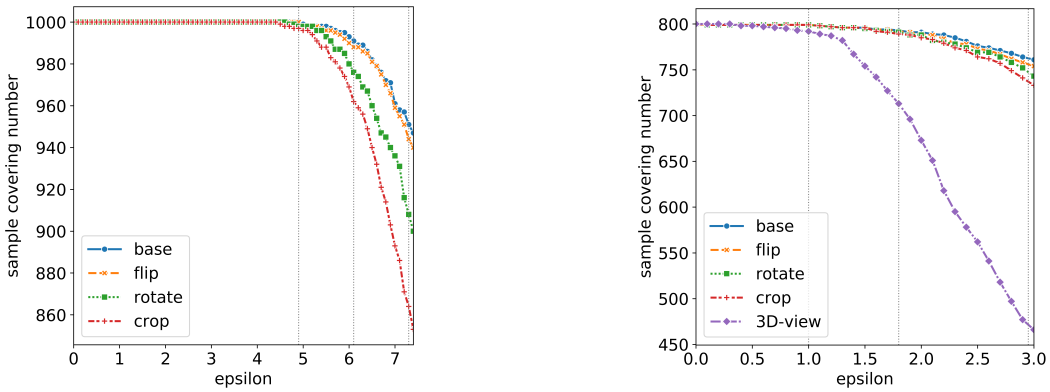
The pool of candidate data transformations on a given dataset may be infinitely large. To maximize the generalization benefit of model invariance, we usually make selections based on expensive cross-validations due to the absence of a model-training-free guidance. Section 3 suggests that invariant models may benefit from improved generalization guarantees if the corresponding data transformations induce small sample covering numbers. Therefore, we propose to use the sample covering number as an empirical suitability measurement to guide the data transformations selection. We discuss its advantages, limitations, and empirical mitigation in this section.

Suitability measurement. To maximize the generalization benefit of model invariance on a dataset $\mathcal{S}$, we measure the suitability of data transformations $\mathcal{G}$ by the sample covering number induced by $\mathcal{G}$ and favor the small ones.

Advantages. One advantage of this suitability measurement is that it is model-training-free. It provides a-priori guidance depending only on the dataset and the data transformations, thus avoids expensive cross-validations and fuels the exploration of new types of data transformations. Another advantage is that it applies to any types of data transformations (including the continuous and non-invertible ones) and provides a uniform benchmark.

Limitations and empirical mitigation. Being model-agnostic also poses two limitations to the suitability measurement. One limitation is that this suitability measurement, while capturing invariant models' reduced generalization gap, ignores their potentially increased empirical error. Note that certain data transformations on a dataset may drastically increase invariant models' empirical error and overturn the benefit of reduced generalization gap. To mitigate this limitation, we consider two necessary conditions for maintaining low empirical error. First, the data transformations should preserve the underlying ground-truth labeling. We may use domain knowledge to meet this condition. Second, the model class should be rich enough to contain a low-error invariant hypothesis. In our experiment, neural networks which are invariant and achieve low training error suffice this condition.

Another limitation is that this suitability measurement ignores models' potential Lipschitz constant change after enforcing the invariance. Theorem 3.4 suggests that the generalization benefit enjoyed by invariant models depends on models' Lipschitz constant and can be overturned if enforcing invariance leads to significantly larger Lipschitz constant. To mitigate this limitation, we use the fact that we are doing classification tasks and use the label information to heuristically offset the Lipschitz constant increase. We use the minimum inter-class distance change after applying data transformations to capture the Lipschitz constant change and use it to normalize the sample covering number for better data transformation selections (see Appendix E.5.2).



(a) CIFAR-10 (b) ShapeNet

Figure 2: Estimated sample covering numbers induced by different data transformations at different resolutions ϵ. "base" indicates no transformation. Note that as ϵ increases, it starts to exceed the L_2 distance between some images and thus some images get covered by others without doing any transformation. Three vertical dashed lines indicate the maximum resolution ϵ at which the "base" yields a certain sample covering number, and from left to right they are $100\%n$, $99\%n$, $95\%n$.

6 Experiments

In this section, we implement the sample cover estimation algorithm and verify the effectiveness of using sample covering numbers to guide the data transformations selection. We first estimate the sample covering number induced by different types of data transformations on some image datasets. Then, we investigate the actual generalization benefit for models invariant to those data transformations and analyze the correlation[4].

Datasets. We report experimental results on CIFAR-10 [27] and ShapeNet [10] in this section, and relegate results on CIFAR-100 and Restricted ImageNet to Appendix E.5.1. ShapeNet is a large-scale 3D data repository which enables us to do more complex data transformations (e.g., change of 3D-view) beyond the common 2D geometric transformations. The work [12] provides 24 multi-view pre-rendered images for each 3D object in 10 chosen categories. For convenience, we use those images to approximate the random perturbation of the 3D-view.

Data transformations. We evaluate some commonly used data transformations with typical parameter settings which we assume to be label-preserving. We choose *flipping*, *cropping*, and *rotation* on CIFAR-10, and additionally consider the *3D-view* change on ShapeNet. We use the same data transformations with the same parameter settings during estimating the sample covering number and evaluating the generalization benefit. Appendix E provides more details of our experimental settings.

6.1 Estimation of Sample Covering Numbers

We implement the algorithm in Section 4 to estimate the sample covering number induced by different data transformations. For efficiency, we randomly sample 1000 training images from CIFAR-10 and randomly sample 800 training images from ShapeNet. Appendix E compares results with different sample sizes. We use the Euclidean norm for defining the sample cover. For continuous data transformations, we do uniform random sampling to approximate the orbit of a data point.

Figure 2 illustrates the estimated sample covering numbers induced by different transformations at different resolution ϵ. As the resolution ϵ increases, the sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by any data transformation starts to decrease, indicating a smaller-sized sample cover needed to cover the entire dataset. Meanwhile, different transformations behave differently. On CIFAR-10, cropping induces the smallest sample covering number. On ShapeNet, 3D-view transformation induces the smallest sample covering number and the gap is significant. Our propositions suggest that data

[4]Code is available at https://github.com/bangann/understanding-invariance.

Model	$n = 100$		$n = 1000$		$n = all$	
	acc (%)	gap	acc (%)	gap	acc (%)	gap
Base	41.05 ± 0.52	58.95 ± 0.52	68.62 ± 0.90	31.38 ± 0.90	85.43 ± 0.35	14.57 ± 0.35
Flip	44.19 ± 0.74	55.81 ± 0.74	75.12 ± 0.20	24.88 ± 0.20	89.67 ± 0.24	10.33 ± 0.24
Rotate	47.02 ± 0.46	52.93 ± 0.51	76.07 ± 0.28	23.92 ± 0.27	89.91 ± 0.13	10.05 ± 0.16
Crop	$\mathbf{50.47 \pm 0.48}$	$\mathbf{49.53 \pm 0.48}$	$\mathbf{81.84 \pm 0.12}$	$\mathbf{18.15 \pm 0.11}$	$\mathbf{92.52 \pm 0.08}$	$\mathbf{7.48 \pm 0.08}$

Table 1: Classification accuracy and generalization gap (the difference between training and test accuracy) for ResNet18 on CIFAR-10. n denotes the sample size per class.

Model	$n = 100$		$n = 1000$		$n = all$	
	acc (%)	gap	acc (%)	gap	acc (%)	gap
Base	67.75 ± 2.02	32.25 ± 2.02	83.33 ± 0.38	16.67 ± 0.38	91.81 ± 0.22	8.18 ± 0.22
Flip	69.75 ± 1.55	30.25 ± 1.55	84.24 ± 0.30	15.76 ± 0.30	92.07 ± 0.20	7.92 ± 0.20
Rotate	70.25 ± 1.19	29.50 ± 1.15	83.93 ± 0.38	15.94 ± 0.35	91.85 ± 0.20	8.03 ± 0.26
Crop	74.88 ± 1.03	23.53 ± 1.30	86.13 ± 0.39	13.75 ± 0.32	92.64 ± 0.12	7.17 ± 0.19
3D-View	$\mathbf{78.13 \pm 1.31}$	$\mathbf{14.94 \pm 1.76}$	$\mathbf{88.79 \pm 0.34}$	$\mathbf{8.38 \pm 0.79}$	$\mathbf{94.38 \pm 0.08}$	$\mathbf{3.09 \pm 0.10}$

Table 2: Classification accuracy and generalization gap (the difference between training and test accuracy) for ResNet18 on ShapeNet. n denotes the sample size per class.

transformations which induce smaller sample covering numbers tends to bring more generalization benefit for the corresponding invariant models. Therefore, Figure 2 indicates that models should generalize well if it is invariant to 3D-view transformation on ShapeNet or to cropping on CIFAR-10.

6.2 Evaluation of Generalization Benefit

We now evaluate the actual generalization performance of invariant models to verify if the sample covering number is a good suitability measurement. We use ResNet18 [24] on both datasets and discuss the influence of model class's implicit bias in Appendix E. A simple method to learn invariant models is to do data augmentation. The augmented loss function is $\mathcal{L}_{aug}(\boldsymbol{x}) = \mathcal{L}(f(g(\boldsymbol{x})))$, where $f(\cdot)$ denotes the model and $g(\boldsymbol{x})$ denotes a randomly sampled example in $\boldsymbol{x}$'s orbit induced by transformation $\mathcal{G}$. We use this method on CIFAR-10 and ShapeNet and show results in Table 1 and 2.

Sample covering number correlates well with generalization benefit. We use the generalization gap (the gap between training accuracy and test accuracy) to measure actual generalization benefit. Compared with the baseline, invariant models show an improved reduced generalization gap and also improved test accuracy. On CIFAR-10, cropping-invariant model shows the smallest generalization gap and the highest accuracy. On ShapeNet, the model that is invariant to 3D-view changes shows the smallest generalization gap and the highest accuracy, especially when the training data size is small. By comparing results in Figure 2 and Table 1-2, we observe a clear correlation between the smaller sample covering number and better generalization benefit. This verifies our proposition — invariance to more suitable data transformations gives the model more generalization benefit.

Model invariance indeed improves after learning. We conduct another experiment on ShapeNet with 3D-view transformation to verify that the improved generalization is indeed brought by model invariance. Instead of using data augmentation, we train models with an invariance regularization loss similar to [47]: $\mathcal{L}_{reg} = \mathcal{L}_{cls} + \mathcal{L}_{inv} = \mathcal{L}(f(\boldsymbol{x})) + \lambda \mathrm{KL}(f(\boldsymbol{x}), f(g(\boldsymbol{x})))$. Specifically, in addition to minimize the classification loss on the original image, we penalize the model by minimizing the KL divergence between model outputs of the original image and the transformed one. By adjusting λ, we can control the invariance extent. At test time, we use an empirical approximation of $\mathcal{L}_{inv}(\boldsymbol{x}) = \mathbb{E}_{g_1, g_2 \in \mathcal{G}}[\mathrm{KL}(f(g_1(\boldsymbol{x})), f(g_2(\boldsymbol{x})))]$ to evaluate model invariance under transformation $\mathcal{G}$. In Table 3, as we increase the invariance penalty by increasing λ, invariant models show smaller generalization gap. Moreover, the decreased invariance loss and increased consistency accuracy verify that the model invariance indeed improves after learning, and thus suggests that the generalization benefit is brought by model invariance.

λ	train acc (%)	test acc (%)	gap	$\mathcal{L}_{inv}$	$\mathcal{A}_{inv}(\%)$
0	99.99 ± 0.01	91.81 ± 0.22	8.19 ± 0.22	0.0548 ± 0.0028	62.0 ± 0.6
0.01	99.98 ± 0.00	92.77 ± 0.16	7.21 ± 0.16	0.0290 ± 0.0029	74.78 ± 1.61
0.1	99.99 ± 0.00	93.87 ± 0.19	6.11 ± 0.19	0.0152 ± 0.0003	83.12 ± 0.50
0.3	99.98 ± 0.00	94.23 ± 0.11	5.76 ± 0.11	0.0121 ± 0.0003	85.10 ± 0.20
1	99.58 ± 0.04	94.68 ± 0.09	4.90 ± 0.09	0.0095 ± 0.0001	86.94 ± 0.08
3	97.74 ± 0.19	94.48 ± 0.19	3.26 ± 0.09	0.0060 ± 0.0003	88.15 ± 0.18
10	95.67 ± 0.26	93.56 ± 0.29	2.11 ± 0.04	0.0037 ± 0.0002	89.20 ± 0.16
100	92.89 ± 0.25	91.85 ± 0.26	1.03 ± 0.03	0.0018 ± 0.0001	89.82 ± 0.10

Table 3: Evaluation of ResNet18 on ShapeNet under 3D-view transformations. $\mathcal{L}_{inv}$ denotes the test invariance loss. $\mathcal{A}_{inv}$ denotes the test consistency accuracy (indicating whether model's prediction is unchanged after data transformation) under worst-case transformations.

7 Related Work

Understandings from the input space perspective. One line of work characterizes the input space of invariant models. [3, 4] show that the invariant representations equivalently reduce the input dimension for downstream tasks and thus significantly reduce the model complexity (exponential in input dimensions) of downstream linear models. [39, 36] essentially factorize the input space into the product of a base space and a finite set of data transformations. Since the covering number needed to cover the base space is smaller, the associated generalization bound for invariant models is reduced. Compared with these works, our work tries to cover the sample instead of the input space which circumvents the strong dependence on input dimensions and also enables practical evaluation.

Understandings from the function space perspective. Another line of work directly characterizes the function space of invariant models. [28] uses PAC-Bayes to show the reduction of generalization upper bound when the model class is symmetrized to be invariant. [19] analyzes the function space under feature averaging operator and shows the first strict generalization gap (instead of upper bound) via a linear model. This line of work, while being elegant, so far restricts the model invariance to be obtained exclusively via feature averaging.

Note that the categorization of different understanding perspectives are only for presentation convenience and hold no formal distinctions. We also mention some work that studies the model invariance but does not focus on understanding the benefit. [1] proves that the VC-dimension of an invariant model cannot be larger than its counterpart. [9] characterizes the general functional representations of invariant probability distributions as well as neural network structures that implement them. [11] uses group theory to show the benefit of learning with data-augmented loss. In addition, [26] shows that model robustness against data transformations can be used as a proxy for predicting models' generalization performance. [8] propose to select data transformations automatically from model training via optimizing a parameterized distributions of data transformations. Interestingly, our sample covering number may be used to determine their regularization coefficients for better trade-offs.

8 Conclusion

In this paper, we study the generalization benefit of model invariance by deriving model complexity bounds based on the sample cover induced by data transformations. We also propose an algorithm to estimate the sample cover and empirically show that the sample covering number can guide the data transformations selection. Hopefully, this work will fuel the exploration of more suitable data transformations on specific datasets. An interesting direction for future work is to consider the implicit bias of model classes to better understand the generalization benefit of model invariance.

Acknowledgements

This work is supported by a startup fund from the Department of Computer Science of the University of Maryland, National Science Foundation IIS CRII Award, DOD-DARPA-Defense Advanced

Research Projects Agency Guaranteeing AI Robustness against Deception (GARD), Air Force Material Command, and Adobe, Capital One and JP Morgan faculty fellowships.

References

[1] Yaser S. Abu-Mostafa. Hints and the VC Dimension. *Neural Comput.*, 5(2):278–288, March 1993. Place: Cambridge, MA, USA Publisher: MIT Press.

[2] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv:1311.4158 [cs]*, Mar 2014. arXiv: 1311.4158.

[3] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised Learning of Invariant Representations in Hierarchical Architectures. *arXiv:1311.4158 [cs]*, March 2014. arXiv: 1311.4158.

[4] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On Invariance and Selectivity in Representation Learning. *arXiv:1503.05938 [cs]*, March 2015. arXiv: 1503.05938.

[5] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks. *arXiv:2004.13617 [cs, stat]*, April 2020. arXiv: 2004.13617.

[6] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[7] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[8] Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, page 17605–17616. Curran Associates, Inc., 2020.

[9] Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *arXiv:1901.06082 [cs, stat]*, September 2020. arXiv: 1901.06082.

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[11] Shuxiao Chen, Edgar Dobriban, and Jane Lee. A Group-Theoretic Framework for Data Augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21321–21333. Curran Associates, Inc., 2020.

[12] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 628–644. Springer, 2016.

[13] Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, June 2016. ISSN: 1938-7228.

[14] Taco S. Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, page 9142–9153, 2019.

[15] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

[16] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[17] R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[18] Shimon Edelman. Class similarity and viewpoint invariance in the recognition of 3d objects. *Biological Cybernetics*, 72(3):207–220, Feb 1995.

[19] Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. *arXiv:2102.10333 [cs, stat]*, February 2021. arXiv: 2102.10333.

[20] Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[21] Robert Gens and Pedro M. Domingos. Deep symmetry networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *NIPS*, page 2537–2545, 2014.

[22] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 297–299. PMLR, 06–09 Jul 2018.

[23] Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, 10(1):1411, Jan 2020.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[25] Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2752–2765. Association for Computational Linguistics, Jul 2020.

[26] Sumukh Aithal K, Dhruva Kashyap, and Natarajan Subramanyam. Robustness to augmentations as a generalization metric, 2021.

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[28] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the Benefits of Invariance in Neural Networks. *arXiv:2005.00178 [cs, stat]*, April 2020. arXiv: 2005.00178.

[29] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[30] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[31] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

[32] Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[33] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[34] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL (1)*, page 5582–5591. Association for Computational Linguistics, 2019.

[35] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[36] Akiyoshi Sannai and Masaaki Imaizumi. Improved Generalization Bound of Group Invariant / Equivariant Deep Networks via Quotient Feature Space. *arXiv:1910.06552 [cs, stat]*, March 2020. arXiv: 1910.06552.

[37] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially Robust Generalization Requires More Data. In *Advances in Neural Information Processing Systems 31*, pages 5014–5026. Curran Associates, Inc., 2018.

[38] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[39] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Generalization Error of Invariant Classifiers. *arXiv:1610.04574 [cs, stat]*, July 2017. arXiv: 1610.04574.

[40] Luke Taylor and Geoff Nitschke. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE, 2018.

[41] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Mądry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[42] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[43] Marysia Winkels and Taco S. Cohen. 3d g-cnns for pulmonary nodule detection. *CoRR*, abs/1804.04656, 2018.

[44] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[45] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, page 7085–7094. PMLR, Jun 2019.

[46] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, page 3391–3401, 2017.

[47] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Learning Representations*, 2019.

[48] Tong Zhang. Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res.*, 2:527–550, Mar 2002.

[49] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NIPS*, page 649–657, 2015.